Reinforcement Learning for Mixing Systems

Optimal Sample Complexity for Discounted and Average Reward Mixing Markov Decision Processes

Shengbo Wang

MS&E — Stanford University

INFORMS 2023

Joint work with Jose Blanchet, Peter Glynn

Outline

1. Introduction
2. Formulation
   - Tabular RL
   - Uniform Ergodicity
3. Literature Review and our Contributions
4. Theoretical Insights: How the Mixing Time Impact Complexity
   - Worst Case Discounted MDPs
   - Mixing Discounted MDPs
5. Appendix
   - Wide Sense Regeneration and Consequences

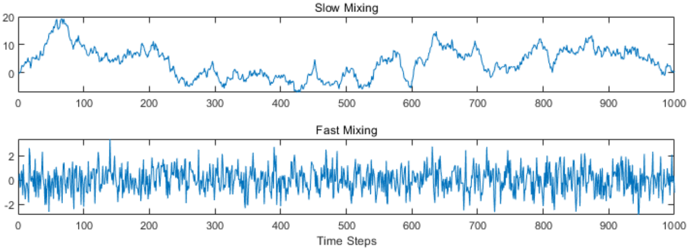
Preliminary Motivation

Dynamic decision making environments in operations research and management science discipline:

- Manufacturing/service networks
- Power grid
- Inventory control
- …

Admissible/optimal (stationary) policies induce mixing: system converge to a unique steady state.

Theoretical Motivation



Rapid mixing Markov chains: Good inference on the steady states can be drawn with less samples.

Sample complexity of RL:

- Estimate the long run average reward using a small sample size.
- Discounted case: effective horizon is long, same can be said.

Outline

1. Introduction
2. Formulation
   - Tabular RL
   - Uniform Ergodicity
3. Literature Review and our Contributions
4. Theoretical Insights: How the Mixing Time Impact Complexity
   - Worst Case Discounted MDPs
   - Mixing Discounted MDPs
5. Appendix
   - Wide Sense Regeneration and Consequences

Tabular RL

Infinite horizon MDP with finite state, action spaces S, A.

- Transition kernel $P = \{p_{s,a} \in \mathcal{P}(S) : (s, a) \in S \times A\}$.
- Suffices to consider *stationary Markov deterministic* policies Π.
- Reward function $\|r\|_\infty \leq 1$.
- Optimal infinite horizon discounted value:

$$v^*(s) = \sup_{\pi \in \Pi} E_s^\pi \sum_{k=0}^{\infty} \gamma^k r(X_k, A_k).$$

- Optimal long run average reward (?):

$$\bar{\alpha} = \sup_{\pi \in \Pi} \lim_{T \to \infty} \frac{1}{T} E_s^\pi \sum_{k=0}^{T-1} r(X_k, A_k)$$

Uniform Ergodicity

Markov kernel induced by $\pi \in \Pi$: $P_\pi(s, s') = p_{s,\pi(s)}(s')$.

A policy $\pi \in \Pi$ is (uniformly) *mixing*: P_π is uniformly ergodic.

Definition (Uniform Ergodicity)

P_π is uniformly ergodic if $\max_{s \in S} \|P_\pi^t(s, \cdot) - \eta_\pi(\cdot)\|_{\mathrm{TV}} \leq c\rho^{-t}$ for all t.

Ways that a MDP can display mixing:

- All policies $\pi \in \Pi$ induces mixing: uniformly ergodic MDP
- Every optimal policy $\pi^* \in \Pi$ is mixing.
- Exists one optimal policy $\pi^* \in \Pi$ that induces mixing.

Uniformly Ergodic MDP

Definition (Uniform Ergodicity)

Uniform ergodicity: $\max_{s\in S} \|P_\pi^t(s,\cdot) - \eta_\pi(\cdot)\|_{\mathrm{TV}} \le c\rho^{-t}$

Mixing time:

$$t_{\mathrm{mix}}(P_\pi) := \inf\left\{t \ge 1 : \max_{s\in S} \left\|P_\pi^t(s,\cdot) - \eta_\pi(\cdot)\right\|_{\mathrm{TV}} \le \frac{1}{2}\right\}.$$

Then $t_{\mathrm{mix}} := \max_{\pi\in\Pi} t_{\mathrm{mix}}(P_\pi) < \infty$ for uniformly ergodic MDP.

Optimal long run average reward:

$$\bar{\alpha} = \sup_{\pi\in\Pi} \lim_{T\to\infty} \frac{1}{T} E_s^\pi \sum_{k=0}^{T-1} r(X_k, A_k)$$

is independent of initial state s.

Outline

Literature Review and our Contributions

Discounted MDPs (v^*). $S, A = O(1)$.

- v^* estimation: optimal error convergence rate for the worst case MDP [Azar et al. 2013; Wainwright 2019]:

$$\epsilon = \widetilde{\Theta}\left(\sqrt{\frac{1}{(1-\gamma)^3 n}}\right) \quad \text{or} \quad n = \widetilde{\Theta}\left(\frac{1}{(1-\gamma)^3\epsilon^2}\right).$$

- The same rate holds for policy learning [Sidford et al. 2018; Agarwal et al. 2020; Li et al. 2022].
- v^* estimation: optimal error convergence rate for mixing MDP [W. et al. 2023a] ($t_{\mathrm{mix}} \le (1-\gamma)^{-1}$; upper and lower bounds):

$$\epsilon = \widetilde{\Theta}\left(\sqrt{\frac{t_{\mathrm{mix}}}{(1-\gamma)^2 n}}\right) \text{ or } \quad n = \widetilde{\Theta}\left(\frac{t_{\mathrm{mix}}}{(1-\gamma)^2\epsilon^2}\right).$$

- The same rate holds for policy learning [W. et al. 2023b].

Literature Review and our Contributions

Average reward MDPs ($\bar{\alpha}$).

Algorithm Idea	Origin	Sample Complexity ($\tilde{O}$)
Primal-dual π learning	[Wang 2017]	$\|S\|\|A\|\tau^2 {t_{\mathrm{mix}}}^2 \epsilon^{-2}$
Primal-dual SMD	[Jin and Sidford 2020]	$\|S\|\|A\| {t_{\mathrm{mix}}}^2 \epsilon^{-2}$
Reduction to DMDP	[Jin and Sidford 2021]	$\|S\|\|A\| t_{\mathrm{mix}} \epsilon^{-3}$
Reduction to DMDP	[W. et al. 2023b]	$\|S\|\|A\| t_{\mathrm{mix}} \epsilon^{-2}$
Lower Bound	[Jin and Sidford 2021]	$\Omega(\|S\|\|A\| t_{\mathrm{mix}} \epsilon^{-2})$

- Contributing literature includes [Wang 2017; Jin and Sidford 2020, 2021; Wang et al. 2022; Zhang and Xie 2023].
- Our algorithm and upper bound in [W. et al. 2023b] settles the optimal policy learning sample complexity!

Comments on the Average Reward Algorithm

[Jin and Sidford 2021]'s reduction approach:

- Reduce the average reward problem to a discounted MDP with long effective horizon $(1-\gamma)^{-1} = \Theta(t_{\text{mix}}\epsilon^{-1})$.
- Use [Li et al. 2022] to solve the discounted MDP.
  Not optimal for mixing MDP!
- The $(1-\gamma)^{-3}$ leads to ϵ^{-3} dependence.

We realize the optimal sample complexity for mixing discounted MDPs. In [W. et al. 2023b]:

- Same reduction $(1-\gamma)^{-1} = \Theta(t_{\text{mix}}\epsilon^{-1})$.
- The algorithm [W. et al. 2023a] requires large initialization sample size.
- Optimize [Li et al. 2022]. Achieve a optimal algorithm for the discounted MDP with small enough initialization sample size.

Outline

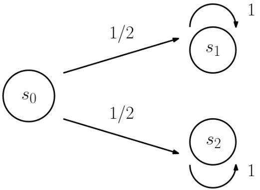
Insights: Worst Case Discounted MDP

Intuition: value function estimation error $\epsilon = \widetilde{O}\left(\sqrt{\frac{1}{(1-\gamma)^3 n}}\right)$.

Consider policy evaluation: estimate v^π for fix $\pi \in \Pi$. Let G denote the realized value. Then in the worst case, the variance is

$$G = \sum_{k=0}^{\infty} \gamma^t r(X_t, A_t); \qquad v^\pi(s) = E_s^\pi G; \qquad \mathrm{Var}_{s_0}^\pi(G) = \Theta\left(\frac{1}{(1-\gamma)^2}\right).$$

Achieved by a chain that is absorbed after one transition, with $r(s_0) = r(s_1) = 0, r(s_2) = 1$.



Insights: Worst Case Discounted MDP Cont'd

Let N be the number of simulations of G. Consider the sample average $\bar{G}_N$. The canonical rate

$$\epsilon \approx \sqrt{\frac{\mathrm{Var}_{s_0}^{\pi}(G)}{N}} \lesssim \sqrt{\frac{1}{(1-\gamma)^2 N}}.$$

Truncate G:

$$\left| \sum_{k=0}^{T} \gamma^t r(X_t, A_t) - \sum_{k=0}^{\infty} \gamma^t r(X_t, A_t) \right| < \epsilon.$$

Then $T = \frac{1}{1-\gamma} \log(\frac{1}{(1-\gamma)\epsilon})$ suffices.

So, $n = TN$, one conjectures that at most

$$\epsilon \le \widetilde{O}\left(\sqrt{\frac{1}{(1-\gamma)^2 n/T}}\right) = \widetilde{O}\left(\sqrt{\frac{1}{(1-\gamma)^3 n}}\right).$$

Key Insights: Mixing Discounted MDP

Non mixing case:

$$\mathrm{Var}_{s_0}^{\pi}\left(G\right) = \Theta\left(\frac{1}{(1-\gamma)^2}\right).$$

Theorem (Variance bound)

If P_π is mixing with mixing time upper bound t_{mix}, then there absolute constant $C > 0$ s.t. $\forall s \in S$ the variance

$$Var_s^{\pi}(G) = Var_s^{\pi}\left(\sum_{k=0}^{\infty} \gamma^t r(X_t, A_t)\right) \leq C \frac{t_{\text{mix}}}{1-\gamma}.$$

This and above insights suggest

$$\epsilon \leq \widetilde{O}\left(\sqrt{\frac{t_{\text{mix}}}{(1-\gamma)} \frac{1}{N}}\right) = \widetilde{O}\left(\sqrt{\frac{t_{\text{mix}}}{(1-\gamma)^2 n}}\right).$$

Thank you for listening!
Your questions and thoughts are most welcome!

[W. et al. 2023a]: Wang, S., Blanchet, J., and Glynn, P. (2023). Optimal Sample Complexity for Average Reward Markov Decision Processes.

[W. et al. 2023b]: Wang, S., Blanchet, J., and Glynn, P. (2023). Optimal Sample Complexity of Reinforcement Learning for Mixing Discounted Markov Decision Processes.

Outline

Wide Sense Regeneration

Theorem (W. et al. 2023)

There exists constants $c, C > 0$ s.t. $ct_{\text{minorize}}(P_\pi) \le t_{\text{mix}}(P_\pi) \le Ct_{\text{minorize}}(P_\pi)$.

Why t_{minorize}? General state space; easier to access.
Split chain $P_\pi^m(s, \cdot) \ge p\psi(\cdot)$:

1. At time t, flip a coin B_{t+m} with success probability p.
2. If $B_{t+m} = 1$, generate $X_{t+m} \sim \psi$; if not, generate $X_{t+m} \sim R(X_t, \cdot)$, $R(s, s') = \frac{1}{1-p}(P_\pi(s, s') - p\psi(s'))$.
3. Generate $X_{t+1}, \ldots, X_{t+m-1}$ condition on X_t, X_{t+m}.

Wide sense regeneration: Let $\tau_{j+1} = \inf\{t > \tau_j : B_t = 1\}$, $W_{j+1} = (X_{\tau_j}, \ldots, X_{\tau_{j+1}-1})$.

- 1-dependent cycles: $\{W_j : 1 \le j \le k\}$ and $\{W_j : k+2 \le j\}$ are independent for all $k \ge 1$.
- Cycles $\{W_j, j \ge 2\}$ are identically distributed.

Wide Sense Regeneration

Additive structure of the value:

$$\sum_{k=0}^{\infty} \gamma^k r_\pi(X_k) = \sum_{j=0}^{\infty} \sum_{k=\tau_j}^{\tau_{j+1}-1} \gamma^k r_\pi(X_k)$$
$$= g_\pi(W_1) + \sum_{j=1}^{\infty} \gamma^{\tau_j} g_\pi(W_{j+1}).$$

Variance computation (suppose $\gamma = 1$, the sum is truncated)

$$\mathrm{Var}\left(\sum_{j=1}^{T} g_\pi(W_{j+1})\right) = \sum_{j=1}^{T} \sum_{k=1}^{T} \mathrm{Cov}\left(g_\pi(W_{j+1}), g_\pi(W_{k+1})\right)$$
$$= T\mathrm{Var}(g_\pi(W_{j+1})) + 2(T-1)\mathrm{Cov}\left(g_\pi(W_2), g_\pi(W_3)\right)$$

by independence and identical distribution.

Variance of Discounted Reward

Variance Bound

If P_π is uniformly ergodic with minorization time $t_{\mathrm{minorize}}(P_\pi)$ and the reward $\|r\|_\infty \leq 1$, then.

$$\mathrm{Var}\left(\sum_{k=0}^{\infty} \gamma^k r_\pi(X_k)\right) \leq c\frac{t_{\mathrm{minorize}}}{1-\gamma}$$

For comparison: worst case variance without mixing: $\Theta((1-\gamma)^{-2})$.